Chapter 1
Entity–Centric Semantic Interoperability

Paolo Bouquet
University of Trento, Italy

Heiko Stoermer
University of Trento, Italy

Wojciech Barczyński
SAP AG, Germany

Stefano Bocconi
Elsevier B.V., The Netherlands

EXECUTIVE SUMMARY

This chapter argues that the notion of identity of and reference to entities (objects, individuals, instances) is fundamental in order to achieve semantic interoperability and integration between different sources of knowledge. The first step in order to integrate different information sources about an entity is to recognize that those sources describe the same entity. Unfortunately, different systems that manage information about entities commonly issue different identifiers for these entities. This makes reference to entities across information systems very complicated or impossible, because there are no means to know how an entity is identified in another system. The authors propose a global, public infrastructure, the Entity Name System (ENS), which enables the creation and re-use of identifiers for entities. This a-priori approach enables systems to reference entities with a globally unique identifier, and makes semantic integration a much easier job. The authors illustrate two enterprise use cases which build on this approach: entity-centric publishing, and entity-centric corporate information management, currently being developed by two leading companies in their respective fields.

INTRODUCTION

The use of semantic techniques for interoperability and integration has been gaining momentum for several years, not to a small extent driven by the efforts in the area of the Semantic Web which, since the beginning of the new millenium, has been occupying scientists and practitioners alike to explore methods that originated from traditional AI,
with the goal of more intelligent and larger-scale information integration.

Substantial efforts have been devoted to effect a similar transition of what the Web achieved with respect to traditional hypertext systems, in the area of semantic representation, integration, and interoperability.

There is however a very important difference between traditional knowledge-based systems and modern approaches that attempt to achieve semantic computing at web scale: the notion of global interlinking of distributed pieces of knowledge.

At the base of such interlinking - and the resulting semantic interoperability of fragments of data - is the notion of identity of and reference to entities. Systems that manage information about entities (such as objects or individuals) commonly issue identifiers for these entities, just in the way relational databases may need to issue surrogate keys to uniquely identify records. If these identifiers are generated by the information system itself, several issues arise that hinder interoperability and integration considerably:

- a proliferation of identifiers is taking place, because the same object is potentially issued with a new identifier in several information systems; therefore, applications need to keep track of a growing amount of identifiers;
- reference to entities across information systems is very complicated or impossible, because there are no means to know how an entity is identified in another system;
- injectivity of identifiers is in general not guaranteed, since the same identifier can denote different entities in different information.

To this end, we propose a global, public infrastructure, the Entity Name System (ENS), which fosters the systematic creation and re-use of identifiers for entities. This a priori approach enables systems to reference the entities which they describe with a globally unique identifier, and thus create pieces of information that are semantically prealigned around those entities. Semantic search engines or integration systems will thus be enabled to aggregate information from distributed sources around entities in a precise and correct way. We call this the entity-centric approach to semantic interoperability.

The ENS is currently under creation in a large European Integrated Project (IP) named OKKAM\(^1\). Part of this project are two enterprise use cases which build on this approach: entity-centric publishing, and entity-centric corporate information management, which are covered by two major companies in their respective fields, and which we are going to describe in detail in this document.

**ENTITY-CENTRIC SEMANTIC INTEROPERABILITY**

Information systems are full of valuable information about entities\(^2\) which are relevant for the business of an organization. This is evidently true for systems based on structured information (like relational databases), but is also true for much less structured types of data, like email folders, text documents, slide presentations, web portals, forums and discussion lists. Being able to collect information about one of these entities, or about their relations with other entities, is a task which would help in many strategic processes, from knowledge management to decision making. However, as we said in the introduction, modern organizations are so complex that there is no centralized control on how and where information is produced and published. This is true not only within the organizational boundaries, but also across organizations and across networks, including of course the Internet. This is why, much more than in the past, the concepts of data interlinking and semantic integration and interop-